Concordance of microarray and RNA-Seq differential gene expression

BF528 Project 3

Yu Zhong, Zhiyu Zhang, Jianfeng Ke, Huisiyu Yu

Introduction

DNA microarray has been a widely adopted tool for gene expression profiling since its invention in the 1990s[1]. Over the past decade, RNA sequencing (RNA-Seq) emerged as an alternative, and gradually became indispensable for transcriptome-wide differential gene expression analyses[2]. Subsequently, the need to evaluate performance and compatibility for these two major platforms also emerged. In their 2014 study, Wang et al. conducted a comparative analysis to evaluate the similarities and differences of RNA-seq and microarrays in terms of differentially expressed genes (DEGs) identification and the accuracy of predictive models generated from these two platforms. Rat liver tissues samples were exposed to varying degrees of perturbation by 27 toxicological treatments that correspond to 7 modes of action (MOA), and sequenced by microarray as well as RNA-Seq in order to assess the influence of chemical (treatment effect) on cross-platform concordance and on the performance of predictive models generated using each technology[3].

In this project we aimed to reproduce part of the concordance analysis of Wang et al's study using a subset of their microarray and RNA-Seq data consisting of three toxicological treatment-control samples, and to compare pathway enrichment results to those reported in the original paper. Our result showed a low overall concordance between the two platforms, and both platforms showed higher concordance for higher treatment effect, with exception to the pirinixic acid treatment group in microarray-based analysis.

Methods

Data Description & Quality Control

3 toxicological treatments (toxgroup_6: 3-methylcholanthrene, fluconazole, pirinixic acid) and their corresponding control samples were selected from the original paper.

Microarray and RNA-Seq data were downloaded from NCBI (National Center for Biotechnology Information) and GEO (Gene Expression Omnibus). RNA-seq reads from FASTQ files were aligned against rat genome using STAR aligner[4]. FastQC[5] and MultiQC [6] were used to inspect the quality of sequencing reads.

Differential Expression Analysis with DESeq2 and Limma

Read counting was performed against reference genome annotation (rn4_refGene) using Subread[7]. Differential expression analyses were performed upon three treatment types with their corresponding control samples on normalized expression matrices for the two sequencing methods (RMA for microarray, normalized counts for RNA-Seq). DEGs were determined using Bioconductor packages DESeq2[8] and limma[9] for RNA-Seq and microarray data respectively, with adjusted P-value <0.05 as threshold for both methods. Up vs down regulation was determined by positive vs negative log2 fold change values respectively (R version 4.0.3). The chemicals used for the three treatment types, their corresponding MOA and abbreviations can be found in **table 2**.

Cross-platform Concordance Analysis

Affymetrix probe IDs were mapped to RefSeq identifiers using a master table created by the original authors. Concordance between DEGs found by RNA-Seq and microarray was calculated for both platforms according to the equations specified in the original paper:
Concordance was computed as:

$$Concordance = 2 \times intersect(DEGs_{microarray}, DEGs_{RNA-Seq})/(DEGs_{microarray} + DEGs_{RNA-Seq})$$

With agreement in directionality of fold change, and adjusted for random overlap between two independent sets. Background corrected intersection was computed as:

$$Background\ corrected\ intersection = (n_0 \times N - n_1 \times n_2)/(n_0 + N - n_1 - n_2)$$

In the above equation, n1 and n2 is the number of DEGs from DESeq2 and limma, respectively; n0 is the raw intersection between the two platforms, and N is the total number of genes detected across both platforms. Concordance analysis and corresponding visualizations were conducted in Python 3.7.9.

Gene set enrichment analysis and Hierarchical clustering analysis

KEGG pathway analysis was performed on differentially expressed genes derived from RNA-seq and microarray results using DAVID Functional Annotation Tool[10]. Pathways under the cutoff of 0.1 were considered as enrichment terms. Taking into account the discrepancy in library size for each sample, we normalized expression counts across samples using standardization method, followed by clustering on those genes with coefficient of variation greater than 0.2.

Results

Alignment Statistics

Results from STAR alignment can be found in **Table 1**.

Table 1. STAR Alignment Statistics

Sample ID	Input reads	Average read length	Uniquely mapped(%)	Multi-mapped(%)	Unmapped(%)
SRR1177963	17897455	202	85.52	3.88	10.60
SRR1177964	19342910	202	85.95	3.89	10.17
SRR1177965	16849678	202	85.71	4.08	10.21
SRR1177997	19746775	202	89.86	4.06	6.08
SRR1177999	21838440	202	89.35	4.18	6.48
SRR1178002	18844950	202	89.71	4.12	6.16
SRR1178014	17524782	100	83.57	7.05	9.37
SRR1178021	17497925	200	82.68	6.14	11.19
SRR1178047	17093302	200	84.52	6.33	9.15

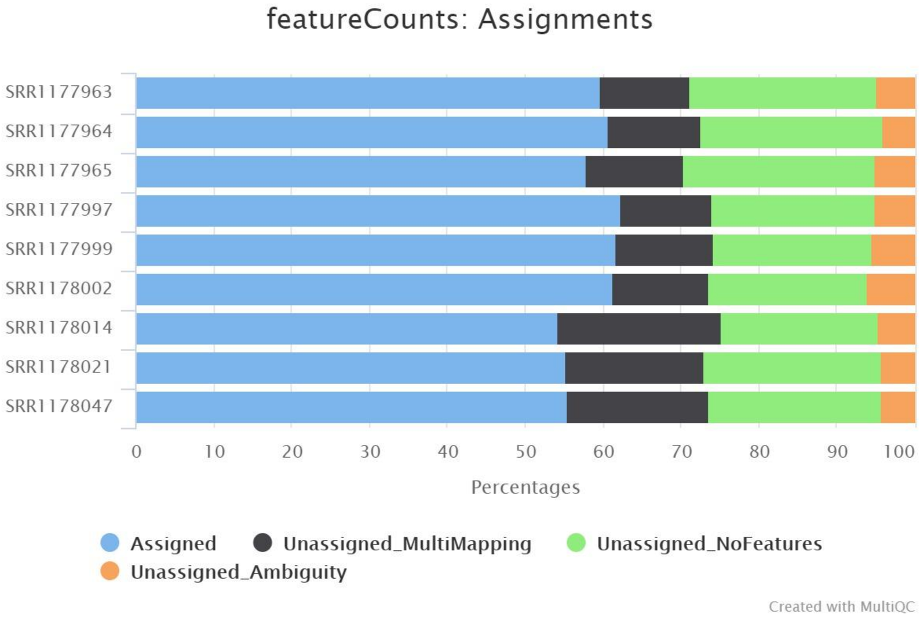

Figure 1. Summary of reads count mapped for genomic features.

To examine the quality of reads mapping to the genome annotation, we summarized a report to highlight that the majority of samples successfully assigned to one feature (**Figure 1**), except for a small proportion with multi-mapping and ambiguity issues. In addition, we plotted gene count distribution across different samples, suggesting that there is no significant difference observed by the overall expression level (**Figure 2**).

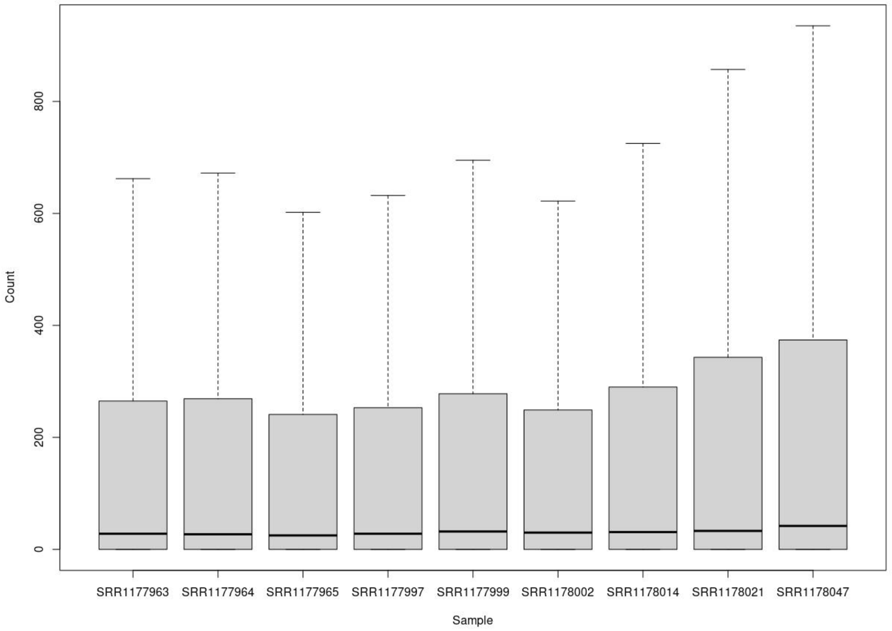

Figure 2. Count distribution across samples.

In order to explore the gene expression profiles within the cells, differential expression analysis was performed on three modes of action for RNA-Seq and microarray data. Number of significantly DE genes are summarized in **Table 2**, together with chemicals used in each treatment and their corresponding MOA. Top ten differentially expressed genes for the 3 MOA identified by both platforms are listed in **Supplementary Table 1**. Histograms of log2 fold change values (**Figure 3**) and volcano plots of log2 fold change vs unadjusted -log10 P-values (**Figure 4**) were used to visually inspect and compare DE results.

Table 2. Chemical names, abbreviations and MOA

Chemical Name	Abbreviation	Modes of Action	# DEGs (RNA-Seq)	# DEGs (microarray)
3-Methylcholanthrene	3ME	AHR	296	58
Fluconazole	FLU	CAR/PXR	3697	1997
Pirinixic Acid	PIR	PPARA	2624	8761

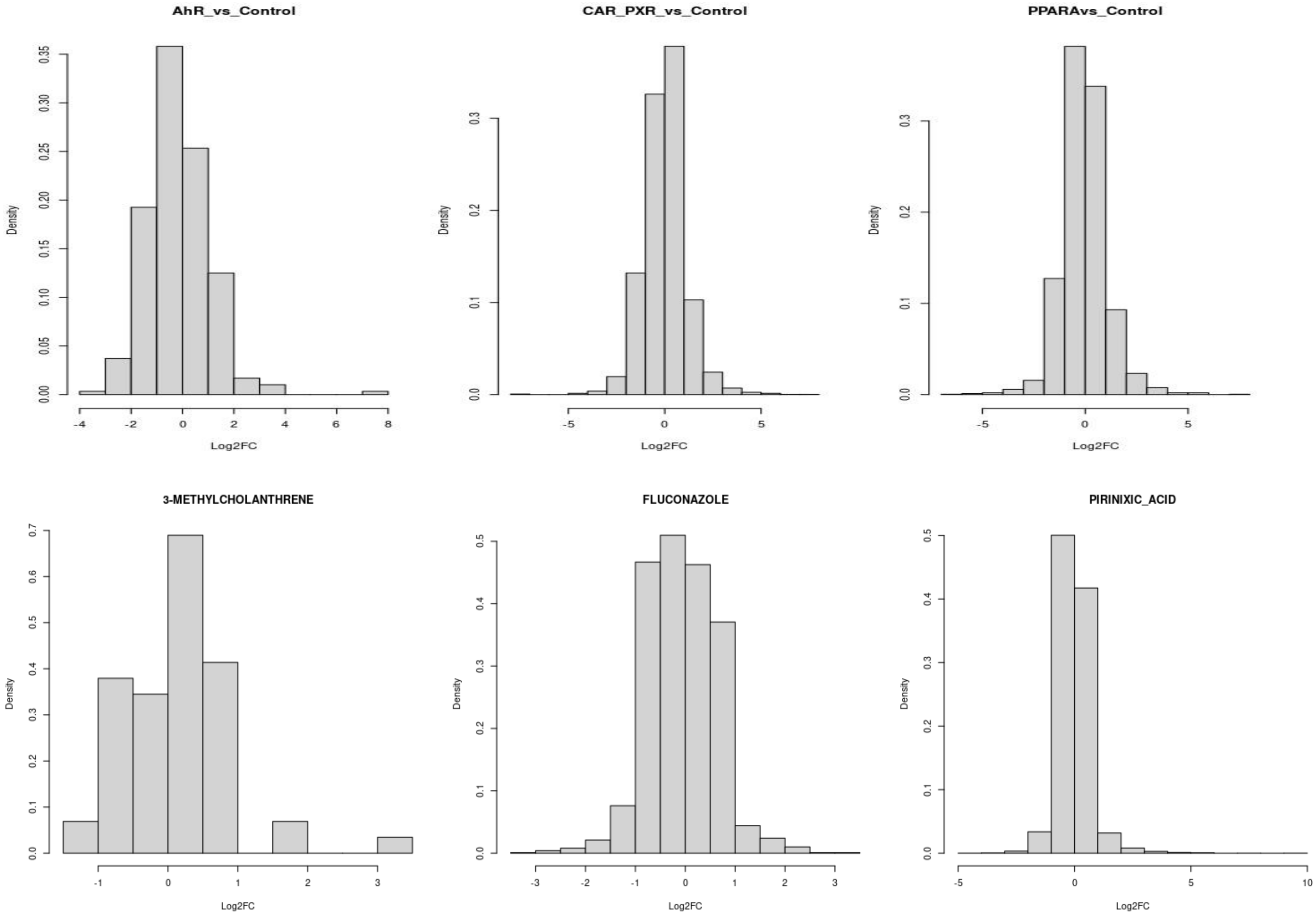

Figure 3. Log2FC distribution over biological comparisons in RNA-seq and Microarray. The top ones are from RNA-seq, and the bottom are from Microarray.

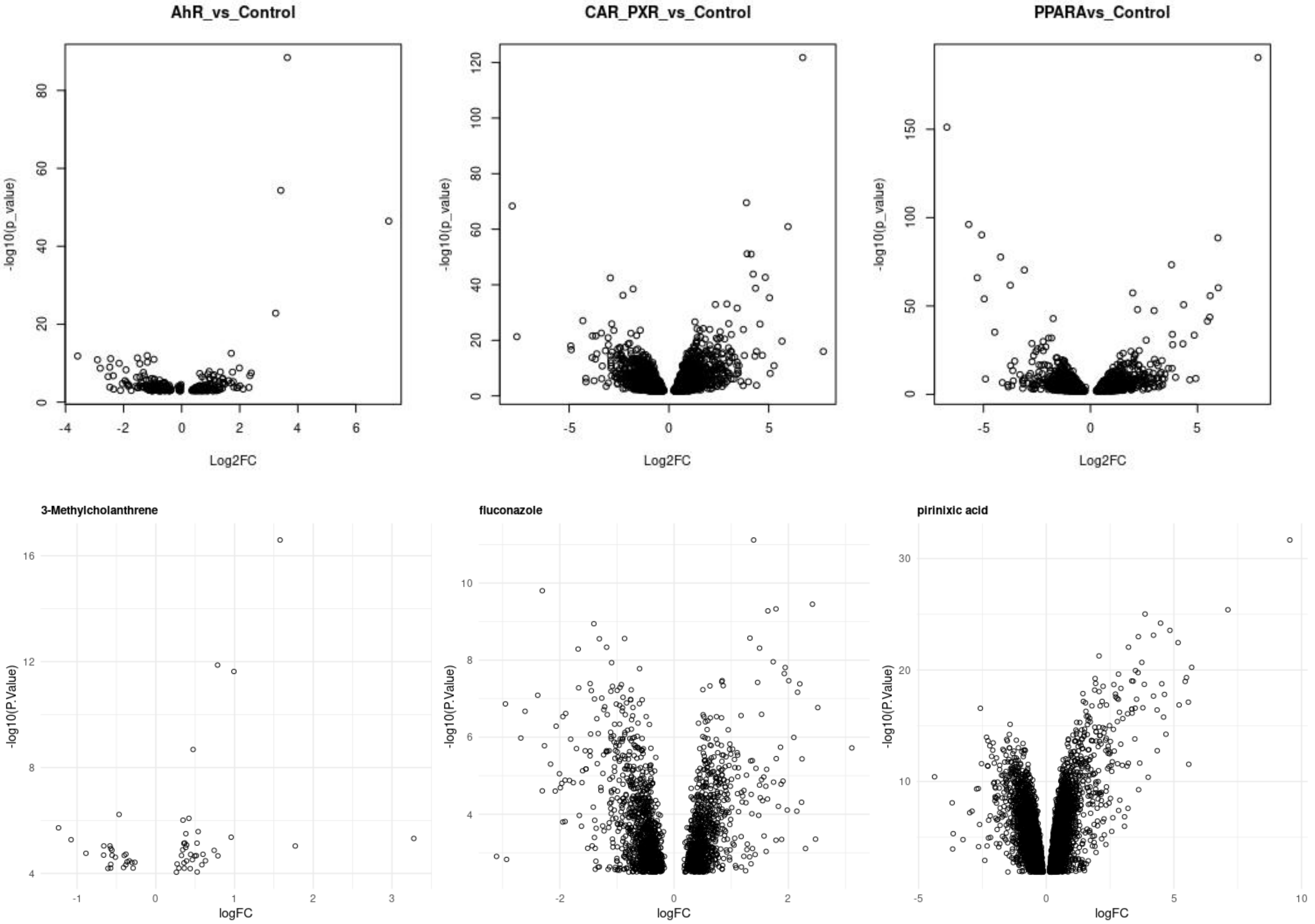

Figure 4. Volcano plot over biological comparisons. The upper ones are from RNA-seq, and the bottom ones are from Microarray. The x-axis indicates log2FC of expression, and the y-axis represents -log10(p value).

Concordance

ID conversion between probe IDs and RefSeq IDs resulted in a considerable amount of unmappable genes. Conversion failure rates are reported in **Table 3**.

Table 3. ID Conversion Failures

	3ME (DE)	FLU (DE)	PIR (DE)	ALL (detected genes)
Probe ID to Refseq ID	12.1%	25.2%	27.2%	40.2%
Refseq ID to Probe ID	9.46%	7.17%	7.01%	9.16%

Due to incomplete mapping between probe IDs and RefSeq IDs, the resulting number of genes detected as well as the number of DE genes after ID conversion were different when the direction of conversion was changed. As a result, concordance appeared to be different when different types of IDs were used (**Figure 5**). Concordance for DEGs between RNA-seq and microarrays were low for all three chemicals examined. 3-Methylcholanthrene had the lowest number of DEGs (**Table 2**) and showed the lowest concordance and highest discrepancy between the two platforms (6.5% using RefSeq IDs, 5.4% using probe IDs). The Fluconazole treatment had the highest concordance relative to other groups (34.3% using RefSeq IDs, 21.7% using probe IDs). With the Pirinixic Acid treatment, the highest number of DEGs were detected, but concordance remained low (14.7% using RefSeq IDs, 8.2% using probe IDs).

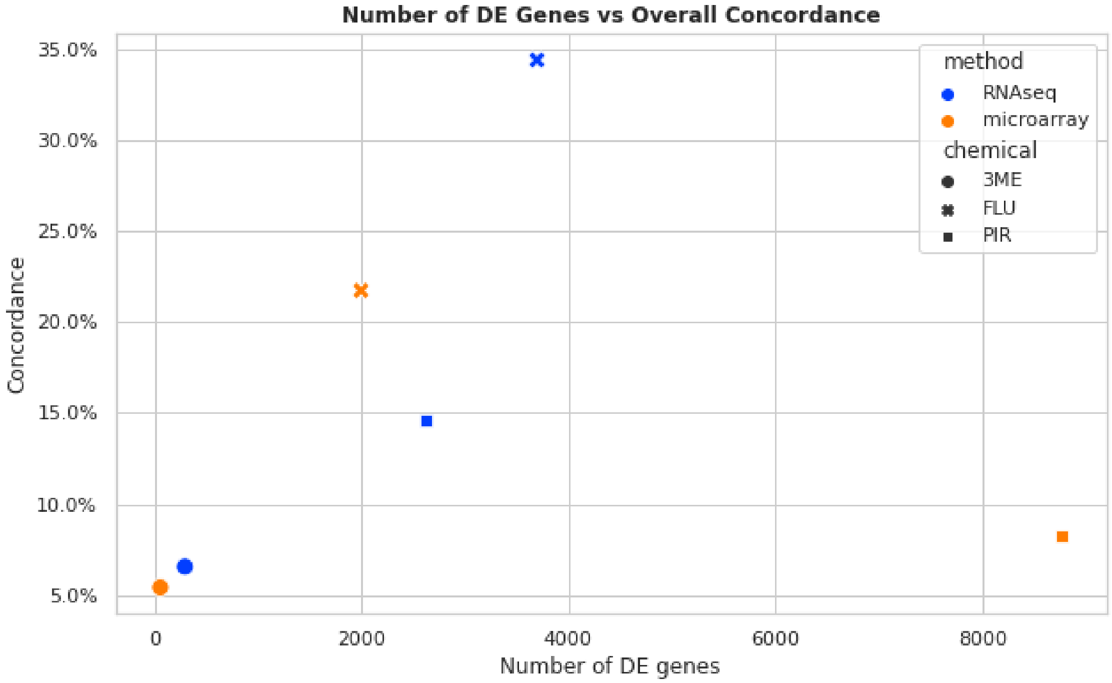

Figure 5. Concordance between RNA-Seq and microarray for 3-Methylcholanthrene (3ME), Fluconazole (FLU), and Pirinixic Acid (PIR). Concordance for significantly DE genes plotted against the number of DE genes identified by DESeq2 (RNA-Seq, blue) and limma (microarray, orange). Method also indicates the type of identifier used as gene IDs.

In order to further examine the effect of expression level on the discrepancy in performance between the two platforms, DEGs were subdivided into above- and below median expression groups for each treatment and platform. The above-median expression group showed consistently higher concordance compared with the below-median expression group (**Figure 6**), with about two fold increase for FLU and PIR. For 3ME, the below median group showed a drastically low concordance between the two platforms (~1% for both types of IDs).

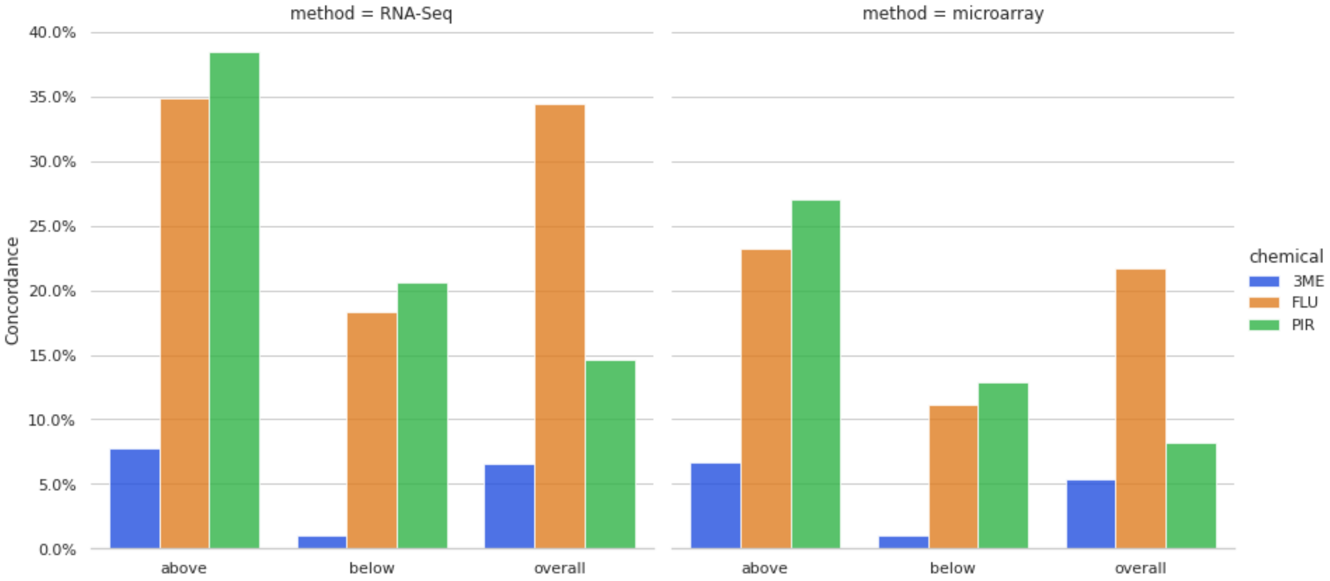

Figure 6. Combined plot of concordance for overall, above-, and below-median expression genes for 3-Methylcholanthrene (3ME), Fluconazole (FLU), and Pirinixic Acid (PIR). Method also indicates the type of identifier used as gene IDs.

In order to evaluate the concordance in enriched biological pathways between microarray and RNA-seq, we performed gene set enrichment analysis for differentially expressed genes derived from the two platforms. We identified a series of enriched pathways for each chemical group that are commonly found in both platforms (**Table 4**), which mainly focused on metabolism related processes.

Table 4. Common pathways enriched for each of the MOA chemical groups shared by both RNA-seq and Microarray platforms

Pathway	P_adj (RNA-seq)	P_adj (Microarray)
PPARA(23)		
Metabolic pathways	4.70E-22	2.20E-27
Complement and coagulation cascades	9.50E-09	7.60E-13
Ribosome	9.50E-09	9.00E-10
Proteasome	4.40E-02	2.50E-09
Fatty acid degradation	2.30E-07	1.50E-08
Biosynthesis of antibiotics	1.50E-06	4.70E-06
Fatty acid metabolism	1.10E-10	4.70E-06
Peroxisome	8.30E-05	6.30E-06
Alzheimer's disease	2.30E-07	2.40E-04
Valine, leucine and isoleucine degradation	8.10E-03	2.40E-04
Oxidative phosphorylation	4.20E-07	2.40E-04
PPAR signaling pathway	7.70E-07	4.20E-04
Fatty acid elongation	3.50E-02	6.20E-04
Staphylococcus aureus infection	1.20E-06	1.00E-03
Pyruvate metabolism	6.10E-02	3.90E-02
Lysosome	4.40E-02	4.40E-02
Pertussis	3.90E-02	4.00E-02
Tryptophan metabolism	3.80E-02	2.70E-02
Ascorbate and aldarate metabolism	2.00E-02	6.00E-02
Propanoate metabolism	8.60E-03	3.80E-02
Biosynthesis of unsaturated fatty acids	3.60E-03	7.60E-03
Carbon metabolism	1.80E-03	2.80E-03
Parkinson's disease	2.20E-07	1.30E-02
CAR/PXR (7)		
Ribosome biogenesis in eukaryotes	2.80E-04	4.80E-07
Metabolic pathways	1.00E-14	3.30E-04
RNA transport	1.20E-02	1.60E-03
Retinol metabolism	6.80E-04	5.60E-03
Ribosome	3.00E-04	1.50E-02

Bile secretion	2.10E-03	2.30E-02
Steroid hormone biosynthesis	4.80E-04	6.80E-02
AhR(4)		
Metabolic pathways	9.10E-05	6.30E-02
Metabolism of xenobiotics by cytochrome P450	2.00E-07	8.90E-02
Retinol metabolism	1.20E-06	9.40E-02
Chemical carcinogenesis	1.20E-08	9.40E-02

To highlight gene expression profiles across different modes of action, we performed unsupervised-clustering analysis among all samples, to demonstrate expression patterns specific to one of chemical treatments. Hierarchical clustering indicates different groups are separated from each other, from which all replicates are well clustered together (**Figure 7**). Also, significant expression patterns may potentially suggest that genes grouped by clusters undergo different biological processes utilizing different modes of action within the cells.

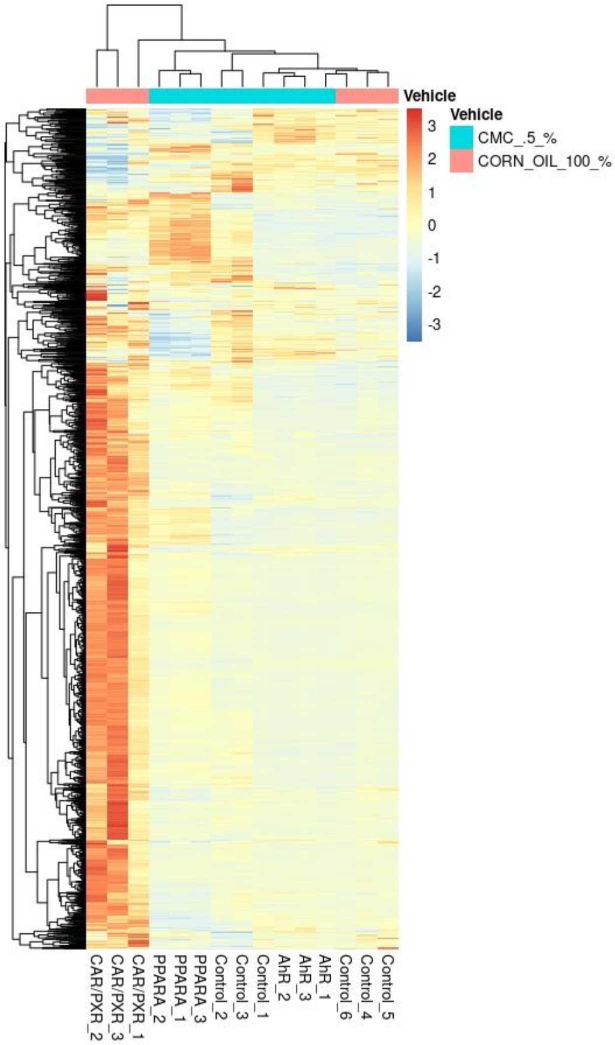

Figure 7. Clustering Analysis of Gene Expression Profiles across different samples.

Discussion

We observed that the majority of reads fall into known genomic features, although a decent fraction of them were unassigned, suggesting that contaminants in library preparation may have contributed to our mapping results. Similiarilly, less adequate quality control of fastq files (Not shown here), can lead biased into downstream analysis.

Mapping probe set IDs to RefSeq IDs resulted in consistently higher rates of failure than mapping from the opposite direction, which was especially evident for all detected genes (40.2%). This could be due to the fact that these genes were identified from a prepared RMA-normalized expression matrix provided for this project, and control probe sets were not removed, therefore remained unmappable. However, all control probe sets were absent from DE analysed genes, and yet a substantial rate of conversion failure still occurred. Since ID conversions were carried out in Python using a prepared conversion matrix, a potential solution could be carrying out this procedure using pre-established Bioconductor packages in R such as gageData instead.

Overall concordance was consistently higher when RefSeq IDs were used, likely an artifact complicated by ID conversion failures. Higher concordance for the above-median expression group was consistent with the findings of Wang et al, indicating that both platforms were better at detecting highly expressed genes. Since only three chemicals were examined, there were not enough data to determine a correlation between treatment effects (represented by the number of DEGs) and cross-platform concordance, but for RNA-Seq, higher number of DEGs did correspond to higher overall concordance, which agrees with the trend observed by the Wang et al. However, for microarray, the PIR treatment group showed the highest number of DEGs (8761) and a low concordance (8.2%), which was inconsistent with the aforementioned trend. This phenomenon could potentially be attributed to lower specificity and accuracy of microarrays in comparison to RNA-Seq, as observed by Wang et al.

Our results of enriched pathways identified by DEGs shared by both RNA-seq and microarray platforms were only partially consistent with those identified in the original paper, namely fatty acid related metabolisms (PPARA) and xenobiotics metabolism (AhR). For CAR/PXR, our result had no overlap with enriched pathways found by Wang et al. This phenomenon might be attributed to the fact that our analysis method did not exclude general housekeeping pathways. As a result, the responses for specific chemical treatments were overwhelmed by highly significantly enriched housekeeping genes. Another possible factor contributing to the discrepancy is that due to unequal number of treatment groups examined, enriched pathways could be inherently different, since the original authors examined three chemicals for each MOA, while we only examined one for each MOA.

Clustering analysis was robust to reveal the biological relationship between samples, and provides us insight into the gene expression patterns within the cells. Interestingly, samples under the condition of CAR/PXR imposed more significant expression profiles than given other mechanisms of action, suggesting a distinct transcriptional response after the treatment.

In conclusion, we found a low overall concordance between RNA-Seq- and microarray-identified DEGs. Concordance tends to be higher for stronger treatment effect, with exception to the pirinixic acid group in microarray-based DE analysis. Higher concordance for above-median expression groups of DEGs suggested better cross-platform compatibility for genes with high expression levels. Most of the challenges encountered in this project pertained to understanding and employing appropriate methodologies. We believe that by adopting alternative methods in our analysis pipelines, such as using different ID conversion strategies and different enrichment analysis tools will improve the reliability of our result.

References

1. 1. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PloS one. 2014;9(1):e78644.
2. 2. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nature reviews. Genetics. 2019 [accessed 2021 Apr 7];20(11):631–656.
3. Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Łabaj, P. P., Kreil, D. P., Megherbi, D., Gaj, S., Caiment, F., van Delft, J., … Tong, W. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nature biotechnology, 32(9), 926–932. https://doi.org/10.1038/nbt.3001
4. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15-21.
5. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.
6. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047-3048.
7. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England), 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656
8. Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.
9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.
10. DAVID: Functional Annotation Tools. Ncifcrf.gov. [accessed 2021 Apr 7]. https://david.ncifcrf.gov/tools.jsp

Supplementary

Supplementary Table 1.
Top ten differentially expressed genes across comparisons in RNA-seq and Microarray.

Gene/Probe ID	Log2FC	P_adj
AhR		

Up-regulated in RNA-seq		
NM_012540	7.132694	1.23E-43
NM_012541	3.639132	3.99E-85
NM_130407	3.414055	2.49E-51
NM_001109459	3.237288	4.06E-20
NM_022521	1.707888	6.51E-10
Down-regulated in RNA-seq		
NM_001109022	-3.582593	2.32E-09
NM_022866	-2.446048	8.62E-09
NM_134329	-1.52241	6.24E-09
NM_053883	-1.186162	2.11E-09
NM_022297	-0.954258	1.17E-08
Up-regulated in Microarray		
1387243_at	1.5784769	7.90E-13
1387759_s_at	0.9926166	2.45E-08
1370613_s_at	0.7846881	2.09E-08
1380888_at	0.5384336	9.08E-03
1383325_at	0.4738773	1.62E-05
1372297_at	0.4205968	4.30E-03
1367669_a_at	0.3819829	9.81E-03
1384544_at	0.3453746	4.32E-03
Down-regulated in Microarray		
1368168_at	-1.2358982	7.37E-03
1387901_at	-0.4668309	3.69E-03
CAR/PXR		
Up-regulated in RNA-seq		
NM_053699	6.70112	1.76E-118
NM_013033	5.9659	3.19E-58
NM_013105	4.82827	2.94E-40
NM_053288	4.33959	2.00E-36
NM_031048	4.21859	2.35E-41
NM_144755	4.11769	1.71E-48
NM_001005384	3.91623	1.54E-48
NM_031605	3.88093	1.55E-66
Down-regulated in RNA-seq		

NM_001130558	-7.85073	1.65E-65
NM_001014166	-2.93597	4.01E-40
Up-regulated in Microarray		
1371076_at	2.4208771	3.28E-06
1390255_at	1.7827581	3.28E-06
1391570_at	1.6411682	3.28E-06
1368731_at	1.3924964	2.37E-07
1380336_at	1.3274522	9.64E-06
Down-regulated in Microarray		
1377014_at	-2.3062528	2.45E-06
1394022_at	-1.4037902	5.89E-06
1398597_at	-1.3073247	9.64E-06
1377192_a_at	-1.1808706	1.34E-05
1372136_at	-0.8666349	9.64E-06
PPARA		
Up-regulated in RNA-seq		
NM_024162	7.82827	2.96E-187
NM_019157	5.95513	6.03E-86
NM_012600	3.78662	7.64E-71
Down-regulated in RNA-seq		
NM_012737	-6.71065	3.69E-148
NM_017158	-5.68982	2.69E-93
NM_001013098	-5.29109	1.28E-63
NM_131903	-5.08478	1.63E-87
NM_001014063	-4.19985	3.72E-75
NM_001013975	-3.74769	1.98E-59
NM_053883	-3.08916	6.34E-68
Up-regulated in Microarray		
1398250_at	9.542023	6.74E-28
1388211_s_at	7.116135	6.12E-22
1389253_at	5.163534	1.37E-19
1375845_at	4.843399	1.72E-20
1387740_at	4.476151	4.85E-21
1374187_at	4.200635	3.93E-20
1391433_at	3.870568	9.83E-22

1386885_at	3.607191	4.52E-20
1384244_at	3.221774	3.09E-19
1367680_at	2.066581	1.70E-18